Additional ways of Integrating Community-based Language Documentation and Language Revitalization

R. Hatcher [1] R. Jimerson [2,3] W. Nephew [3]
M. Jones [3] J. Cordani [1] L. Cremean [1]
E. Prud'Hommeaux [2,4]

[1]University at Buffalo
[2]Rochester Institute of Technology
[3]Seneca Nation of Indians
[4]Boston College

February 28, 2019

The Project
Deep learning speech recognition for Seneca (see)

Goal: Utilize 'cutting-edge' computational techniques to develop tools for revitalizing and documenting an under-resourced language

The Project
Deep learning speech recognition for Seneca (see)

Goal: Utilize 'cutting-edge' computational techniques to develop tools for revitalizing and documenting an under-resourced language

Tools:

- Automatic Speech Recognition (cf. SIRI)

The Project
Deep learning speech recognition for Seneca (see)

Goal: Utilize 'cutting-edge' computational techniques to develop tools for revitalizing and documenting an under-resourced language

Tools:

- Automatic Speech Recognition (cf. SIRI)
- Machine translation from Mohawk to Seneca

Seneca Language & Community

Seneca (Onödowa'ga:) is westernmost member of Haudenosaunee Confederacy.

Seneca Language & Community

Seneca (Onödowa'ga:) is westernmost member of Haudenosaunee Confederacy.

- Appx. 8000 members mostly residing in or around the Tonawanda, Cattaraugus, and Allegany reservations

Seneca Language & Community

Seneca (Onödowa'ga:) is westernmost member of Haudenosaunee Confederacy.

- Appx. 8000 members mostly residing in or around the Tonawanda, Cattaraugus, and Allegany reservations
- Today fewer than 50 L1 speakers remaining

Seneca Language & Community

Seneca (Onödowa'ga:) is westernmost member of Haudenosaunee Confederacy.

- Appx. 8000 members mostly residing in or around the Tonawanda, Cattaraugus, and Allegany reservations
- Today fewer than 50 L1 speakers remaining
- Categorized as shifting (level 7) on EGIDS (Lewis, Simons & Fennig 2015), but probably 8a is more accurate.

Seneca Language & Community

Seneca (Onödowa'ga:) is westernmost member of Haudenosaunee Confederacy.

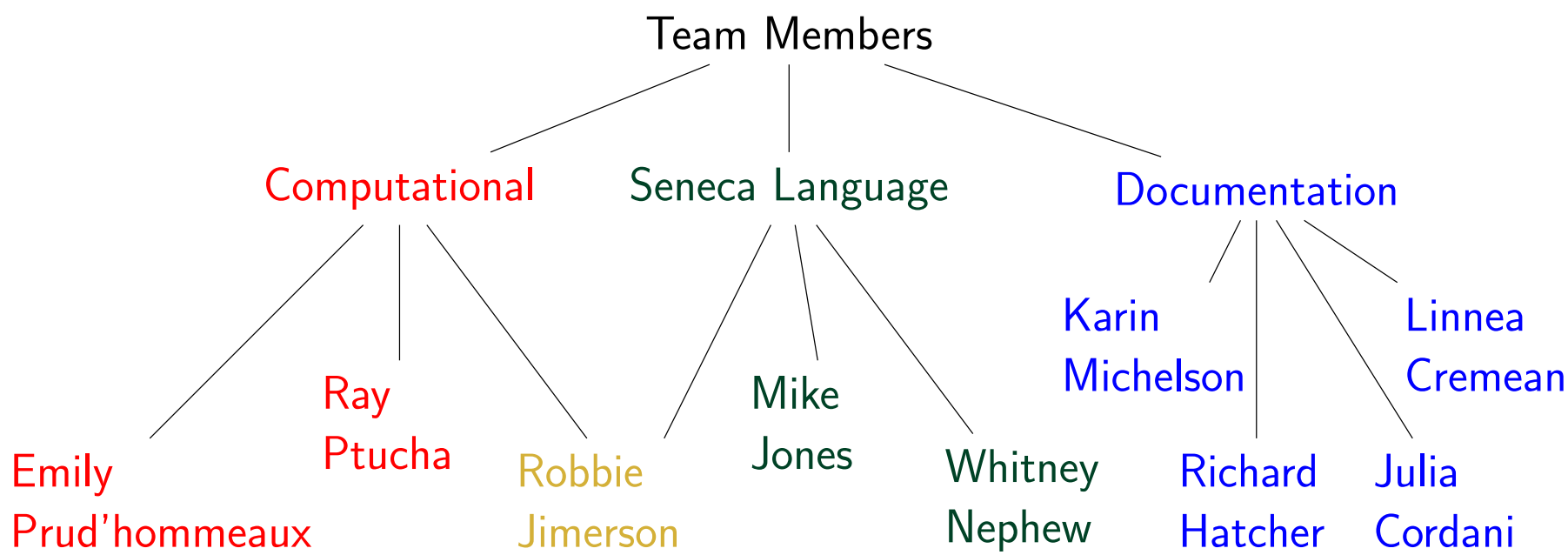
- Appx. 8000 members mostly residing in or around the Tonawanda, Cattaraugus, and Allegany reservations
- Today fewer than 50 L1 speakers remaining
- Categorized as shifting (level 7) on EGIDS (Lewis, Simons & Fennig 2015), but probably 8a is more accurate.

However, there is a active growing community of L2 language speakers.

- Speak with first langauge speakers daily
- Numerous adult immersion programs

Seneca Territory - c. 1650 & present

Team Members



Earlier models of Research

The Haudenosaunee peoples have a long history of being researched by linguists and anthropologists.

- Lounsbury's (1946, 1953) work on Oneida language
- Fenton's (1936, 1942) studies on Seneca ceremony
- Chafe's (1963, 1967, 1977) grammatical work on Seneca

"These peoples of the Longhouse are among the most studied of all North America Indians; indeed, many would argue that they are *over*studied." (Richter 1992)

Community-based Research

Recent trend in linguistic fieldwork particularly in North America & Australia

- "...[R]esearch that is on a language, and that is conducted *for*, *with*, and *by* the language-speaking community within which the research takes place and which it affects." (Czaykowska-Higgins 2009)
- "...community involvement through all stages of the research." (Rice 2011)

Computational Community-based Research

CoLang 2018 - Learning to collaborate

Benefit from attending CoLang together!
Before CoLang

- Thought of collaboration as 'interview process' convincing community of linguist's value
- Skeptical of working with linguist

After attending CoLang

- Importance of developing relationships and honest in the motivations and goals
- Importance of practical skills, e.g. ELAN, Audacity, FLEx, equipment choices & archiving practices

Main Goal of Project

- Speech data collection for Seneca ASR development
  - previously recorded speech (by linguists, community)
  - contemporary recording
- Ancillary uses of ASR
  - Automated first-pass for documentary transcription
  - Force aligned natural speech corpus for corpus phonological studies

Secondary Goals

- FLEx repository for L2 teachers
  - Notable degree of linguistic variation within Seneca speech community cf. (Goddard 2010, Bird 2008)
  - Most language teachers are L2 learner over 70 years
  - Students can question the language judgments and curriculum decisions
  - A repository of annotated Seneca speech data, easily accessible by L2 Seneca language teachers
    - aid in curriculum design
    - provide "justification" for contradictory judgments
- Develop machine translation tools for converting Mohawk materials into Seneca

Secondary Goals

- FLEx repository for L2 teachers
  - Notable degree of linguistic variation within Seneca speech community cf. (Goddard 2010, Bird 2008)
  - Most language teachers are L2 learner over 70 years
  - Students can question the language judgments and curriculum decisions
  - A repository of annotated Seneca speech data, easily accessible by L2 Seneca language teachers
    - aid in curriculum design
    - provide "justification" for contradictory judgments
- Develop machine translation tools for converting Mohawk materials into Seneca
  - Seneca Gospels - published 1878 - using orthography developed by Asher Wright

Secondary Goals

- FLEx repository for L2 teachers
  - Notable degree of linguistic variation within Seneca speech community cf. (Goddard 2010, Bird 2008)
  - Most language teachers are L2 learner over 70 years
  - Students can question the language judgments and curriculum decisions
  - A repository of annotated Seneca speech data, easily accessible by L2 Seneca language teachers
    - aid in curriculum design
    - provide "justification" for contradictory judgments
- Develop machine translation tools for converting Mohawk materials into Seneca
  - Seneca Gospels - published 1878 - using orthography developed by Asher Wright
  - Create deep learning Seq2Seq translation model

Relationship between Documentation and Revitalization

Documentation feeds revitalization with linguistic material

- Recordings
- Sketch/pedagogical grammar
- Dictionary

Revitalization can feed documentation

- Documentation tools
- Recording opportunities
- Corpora (spoken, text)

Acknowledgements

We would like to give thanks to all our elder speakers across the Seneca communities.

- esp. Dr. Hazel Dean & Sandy Jimerson-Dowdy

Support from NSF RI/DEL grant 1761562 & 1761477

- *Collaborative Research: Deep learning speech recognition for documenting Seneca, a Native American language, and other acutely under-resourced languages*

Bird, Sonya. 2008. An Exemplar Dynamic approach to language shift. *Canadian Journal of Linguistics/Revue canadienne de linguistique* 53(2-3). 387–398.

Chafe, Wallace. 1963. *Handbook of the Seneca language*. New York State Museum and Science Service.

Chafe, Wallace. 1967. *Seneca Morphology and Dictionary*. Smithsonian Contributions to Anthropology.

Chafe, Wallace. 1977. Accent and related phenomena in the Five Nations Iroquois languages. In Larry Hyman (ed.), *Studies in stress and accent*, vol. 4, 169–181. Southern California Occasional Papers in Linguistics.

Czaykowska-Higgins, Ewa. 2009. Research models, community engagement, and linguistic fieldwork: Reflections on working within Canadian indigenous communities. *Language documentation & conservation* 3(1). 15–50.

Fenton, William N. 1936. *An outline of Seneca ceremonies at Coldspring longhouse*. Section of anthropology, Department of the social sciences, Yale university.

Fenton, William N. 1942. *Songs from the Iroquois Longhouse: Program Notes for an Album of American Indian Music from the Eastern Woodlands...(From Records in the Archive of American Folk Song, the Library of Congress)*. Smithsonian Institution.

Goddard, Ives. 2010. Linguistic variation in a small speech community: the personal dialects of Moraviantown Delaware. *Anthropological Linguistics* 52(1). 1–48.

Lewis, M. Paul, Gary F. Simons & Charles D. Fennig (eds.). 2015. *Ethnologue: languages of the world, eighteenth edition*. Available online at http://www.ethnologue.com/. Dallas, Texas: SIL International.

Lounsbury, Floyd G. 1946. *Phonology of the oneida language*. University of Wisconsin–Madison MA thesis.

Lounsbury, Floyd G. 1953. *Oneida Verb Morphology*. Yale University Press.

Rice, Keren. 2011. Documentary linguistics and community relations. *Language Documentation & Conservation* 5. 187–207.

Richter, Daniel K. 1992. *The ordeal of the longhouse: The peoples of the Iroquois League in the era of European colonization*. UNC Press Books.